采用 URL 特征的 Hub 网页识别方法研究*

张　策[1]　都云程[1,2]　梁　然[2]
[1](北京信息科技大学 TRS 软件开放实验室　北京 100085)
[2](北京拓尔思信息技术股份有限公司　北京 100101)

摘要:【目的】通过构建简单数据样本, 解决传统网页类型识别方法效率低的难题。【方法】采用 URL 特征作为识别依据, 抽取 URL 信息构建训练集与测试集, 使用支持向量机(SVM)建立机器学习模型以提高识别效率。【结果】在同样的数据集上,该方法的准确率为 91.2%, 优于其他识别方法。在效率性能方面, 该方法提升近 60%。【局限】当遇到 URL 特征不明显甚至完全相背的网站时, 识别准确率会大幅度降低。【结论】该方法在效率方面存在很大优势, 应用到采集系统中可提高采集效率。

关键词: URL 特征　Hub 网页　支持向量机
分类号: TP391.1　G35

1　引　言

随着网络的发展, Web 上的网页数量增长迅猛, 即使采用大规模的分布式网页采集系统, 采集整个网络中的绝大多数重要网页也要花费很长时间。研究结果表明, 中国的网页一个月内大约只有 8.52%发生变化[1], 所以采用全采集的方式, 存在很大的资源浪费。另外由于两次采集的周期过长, 在此周期内网页变化频率大的网页发生了多次变化, 而采集系统不能及时抓取变化后的网页, 就会导致搜索引擎系统不能对这些网页提供检索服务。为了解决这个问题, 产生了网页增量采集系统。

网页增量采集系统不是采集所有得到的 URL, 只是通过估计网页的变化规律采集新出现的网页、变化的网页和消失的网页, 不关心没有变化的网页。这样极大减少了采集量, 能快速同步 Web 上的网页与搜索引擎中的网页, 从而给用户提供更实时的检索服务。

在增量式采集研究中, 网页通常被分为目录型网页(Hub 网页)与主题型网页(Topic 网页)[2], Hub 网页在网站中的作用是引导用户找到相关的主题网页, 相当于目录索引, 没有具体表达的内容, 为主题型网页提供入口[3]。主题型网页是具体讲述某一主题。经实验证明, 很多新网页都是从 Hub 网页链接过去的[4]。因此, 增量式采集系统只要找出 Hub 网页进行采集就能发现新出现的 URL。如上所述, 识别哪些网页是 Hub 网页就成为首先要解决的问题。

针对此问题, 本文提出一种基于 URL 特征的 Hub 网页识别方法, 首次将 URL 特征作为 Hub 网页识别的全部依据, 这将会弥补传统 Hub 网页识别所带来的巨额开销, 最后通过对比实验验证该方法的有效性。

2　相关工作

目前主要的 Hub 网页识别方法有基于简单规则的识别方法[4]、基于多特征启发式规则的分类方法[5-6]和基于网页内容的机器学习方法[7-9]。

基于简单规则的识别方法是分析 Hub 网页 URL 的特点, 总结出其规律, 制定简单规则, 符合条件的就是 Hub 网页。Meng 等提出选择网站首页, 以及网站中网页文件名包含 index、class 和 default 等单词的网页作为 Hub 网页[4], 采集 Hub 网页中链接所对应的网页。该方法能采集到一大部分新网页, 但是对新网页

通讯作者: 张策, ORCID: 0000-0001-6640-4460, E-mail: smiling_boy@163.com。
*本文系国家自然科学基金项目“网页内容真实性评价研究”(项目编号: 61171159)的研究成果之一。

采集的召回率不是很高。存在以下问题:

(1) Hub 网页选择不准确。由于网页的文件名是由人命名的, 没有固定模式, 因此不可能寻找到一个规则可以正确找出所有 Hub 网页;

(2) 不能自动识别 Hub 网页。由于在采集过程中不能及时发现新的 Hub 网页, 所以就不能发现新 Hub 网页中的链接信息。

为了解决基于简单规则方法的局限性, Ail 等提出基于多特征启发式规则的网页分类方法, 依据非链接字符数、标点符号数和文字链接比三个特征构建启发式规则[5]。研究发现 Hub 网页与主题网页在这些特征值上存在广泛差异, 这种差异证明了网页通过这些特征值进行分类的可行性。该方法通过统计网页中各个特征的具体值, 根据贝叶斯公式计算各个特征值对 Hub 网页的概率支持度, 根据每个特征值的概率支持求出综合支持度, 通过与设定的阈值进行比较, 判断网页属于哪一类。该方法的不足之处在于过度依赖阈值的设定, 阈值的设定会直接影响分类的准确率, 然而对于不同类型网站, 阈值设定也不同, 这样就增加了算法的复杂度。

为了解决阈值的依赖问题, 文献[9]提出基于网页内容的机器学习方法, 通过 HTML 解析分析出网页特征, 建立训练集与测试集, 从而得到机器学习模型, 用于识别 Hub 网页。该方法准确率高, 但是效率不高, 增加了系统的额外开销。因为该方法是建立在网页内容的基础上, 需要解析所有的 HTML 网页, 并提取其中的特征进行保存, 这样就在一定程度上占用了系统资源, 给采集系统带来额外负担, 影响采集系统的性能。

上述方法从不同层面对识别 Hub 网页进行分析, 在前人研究的基础上, 本文提出的基于 URL 特征的识别方法, 将会很大程度地解决上述问题。该方法采用 URL 特征作为样本, 选用 SVM 作为机器学习方法进行识别。与基于规则和基于网页内容的方法相比, 提供了一种更具使用价值的方法。一方面, 特征提取简单高效, 易于实现, 同时兼顾了识别的准确率。另一方面, 在采集系统中, 从网页中提取 URL 是必不可少的部分, 因此选用 URL 作为识别依据, 可以减小对系统效率的影响, 不会给采集系统增加太大的额外开销。

3　基于 URL 特征的 Hub 网页识别方法

3.1　SVM 介绍

支持向量机(Support Vector Machines, SVM)是由 Vapnik 等开发的一种机器学习方法。支持向量机是建立在统计学理论——VC 维理论和结构风险最小原理基础上的, 特别是在样本数目较少的情况下, SVM 的性能明显优于其他算法[10-12]。

其基本思想为: 定义最优线性超平面, 将寻找最优超平面的算法归结为求解一个最优(凸规划)问题。进而基于 Mercer 核展开定理, 通过非线性映射, 将样本空间映射到一个高维乃至于无穷维的特征空间, 使在特征空间可以应用线性学习机的方法解决样本空间中高度非线性分类和回归等问题。其还包括以下优点:

(1) 基于结构风险最小化原则, 这样可以避免过拟合问题, 泛化能力强。

(2) SVM 有坚实理论基础的小样本学习方法。它基本上不涉及概率测度及大数定律。从本质上看, 避开了从归纳到演绎的传统过程, 实现了高效的从训练样本到预测样本的“转导推理”, 大大简化了通常的分类和回归等问题。

(3) SVM 的最终决策函数只由少数的支持向量所确定, 计算的复杂性取决于支持向量的数目, 而不是样本空间的维数, 这在某种意义上避免了“维数灾难”。

(4) 少数支持向量决定了最终结果, 这有助于抓住关键样本、“剔除”大量冗余样本, 而且注定了该方法算法简单, 同时具有较好的“鲁棒”性。

3.2　方法概述

Hub 网页识别可以理解为一个二分类问题, 其中正类为 Hub 网页, 负类为主题网页, Hub 网页识别的关键是如何正确划分 Hub 网页与主题网页。

基于 URL 特征的 Hub 网页识别方法主要依据 URL 中与 Hub 网页有关系的特征划分网页。具体过程如下: 分析已经得到的 URL, 提取其中包含的特征信息, 找出与 Hub 网页有关的特征; 将得到的特征整合成训练集与测试集, 用训练集去训练 SVM 机器学习模型, 同时评测其效果; 根据效果调整 SVM 模型参数, 从而确定最优参数, 得到最终 SVM 学习模型。

3.3　实现流程

图 1 展示了基于 URL 特征的 Hub 网页识别方法的架构, 从整体角度出发, 该方法主要包含三大模块:

预处理、特征提取、训练分类。

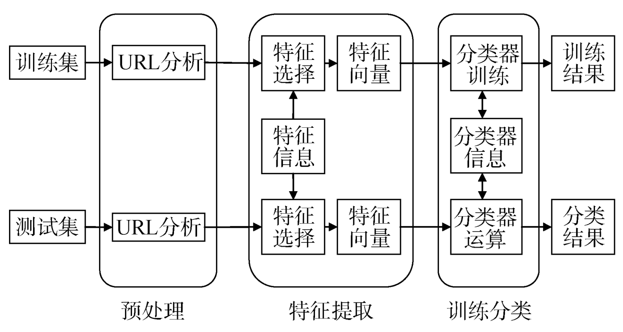


图 1 Hub 网页识别的架构

(1) 预处理

预处理主要包括 URL 分析。URL 中包含很多信息，其中一些信息可以作为网页分类的依据，URL 分析的目的在于找出对分类有用的特征信息。URL 中存在的信息包括 URL 长度以及 URL 之中是否包含某些字符串等。URL 所对应的锚文本也能在一定程度上反映网页类型，因此，需要在预处理阶段提取 URL 所对应的锚文本。

本实验的基础数据都是预先由网络采集器采集，在采集过程中，URL 及其对应的标题等采集信息会作为日志文件被记录下来。因此，实验通过抽取日志文件内容进行分析，获取其中 URL 相关信息。其中包括 URL 标题长度、URL 长度、URL 是否含有日期、网页文件名、文件类型、参数名称、参数个数、目录名称、目录深度、URL 大小、采集深度。

(2) 特征提取

特征提取主要包括特征选择与特征量化。特征选择的任务是要将信息量小、不重要的特征从特征项空间中删除，从而降低特征项空间的维数。特征量化是将选择的特征进行数值化，从而代表该特征和 Hub 页的关联程度。

经 URL 分析，可以得到 URL 中包含的信息，通过查阅相关文献并观察统计可以发现 Hub 网页具有以下区别于主题网页的特征:

①URL 标题长度: 即锚文本长度，Hub 网页由于不讲述某一具体内容，锚文本长度一般较短。

②URL 长度: 由于 Hub 网页基本都位于主题网页上层，因此 Hub 网页的 URL 相比主题网页较短。

③URL 是否含有日期: 主题网页主要讲述某一内容，在 URL 中大多包含发布日期，Hub 网页基本没有。

④网页文件名: Hub 网页 URL 一般有两种可能: 只是一个目录，不存在文件名; 文件名中大都包含“index”、“class”等词语。

⑤文件类型: 文件类型与网页文件名是一体的，存在网页文件名的 Hub 网页大多数是 ASP、JSP、ASPX 和 PHP 类型。

⑥参数名称: 存在参数的 URL 中，主题网页 URL 大都包含 ID 参数，而 Hub 网页的 URL 一般没有。

⑦参数个数: Hub 网页 URL 大多没有参数。

⑧目录深度: Hub 网页基本上都是位于网站的上层。

⑨URL 大小: 即 URL 对应网页的大小。Hub 网页存在大量链接，对应网页也相对较大。

⑩采集深度: 采集到该 URL 的层次。Hub 网页为主题网页提供链接入口，因此，Hub 网页采集一般都先于主题网页。

机器学习模型只能将数值类型进行分类，因此需要将文本类型进行数值化，数值化的依据为归纳不同类型 URL 的文本值，找出代表性的文本值进行赋值，赋值是通过统计求出各个文本值的出现频率，然后计算其出现概率并进行归一化处理。在统计中，选取 500 个 Hub 网页，对各个文本值数量进行统计并计算概率，将概率乘以 100 进行赋值(只是为了让最后得到的特征值在一个合理的范围)，具体过程如下:

①网页文件名为“空”的数量为 302，其概率为 0.604，赋值为 60.4; 含有“class”、“index”、“default”和“list”的数量为 153，其概率为 0.306，赋值为 30.6; 含有“article”和“content”的数量为 0，其概率为 0，赋值为 0; 其他情况数量为 45，其概率为 0.09，赋值为 9。

②文件类型为“空”的数量为 302，其概率为 0.604，赋值为 60.4; 含有“asp”、“jsp”、“aspx”和“php”的数量为 123，其概率为 0.246，赋值为 24.6; 含有“shtml”、“html”和“htm”的数量为 75，其概率为 0.15，赋值为 15; 其他情况数量为 0，其概率为 0，赋值为 0。

③参数名称为“空”的数量为 412，其概率为 0.824，赋值为 82.4; 含有“id”的数量为 52，其概率为 0.104，赋值为 10.4; 其他情况的数量为 36，其概率为 0.072，赋值为 7.2。

(3) 训练分类

通过以上步骤，将 URL 表示成向量空间，使用 LibSVM[13]对 URL 进行分类。LibSVM 是一个快速有效的 SVM 模式识别与回归的集成包，还提供了源码，可以根据需求对源码进行修改。本实验使用的是 LibSVM-3.20 版本①中的 Java 源代码，在参数设置和训练模型两个方面对源码进行修改，增加参数自动寻优

①http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

以及模型文件返回保存功能。

①按照 LibSVM 所要求的格式准备数据集。

该算法使用的训练数据和测试数据文件格式如下:

[label] [index1]:[value1] [index2]:[value2]…

[label] [index1]:[value1] [index2]:[value2]…

其中, label(或称 class)是本条数据所属种类, 通常是一些整数; index 表示特征的序号, 通常是以 1 开始的整数; value 是特征值, 通常是一些实数。当特征值为 0 时, 特征序号和特征值都可以省略, 所以 index 可以是不连续的自然数。

②对数据进行简单的缩放操作。

扫描数据, 由于原始数据可能范围过大或过小, svmscale 可以先将数据重新缩放到适当的范围, 默认范围是[–1,1], 可以用参数 lower 和 upper 分别调整缩放的上界与下界。这样也避免在训练时为了计算核函数而计算内积的时候引起数值计算的困难。

③选用 RBF 核函数。

SVM 的类型选择 C-SVC, 即 C 类支持向量分类机, 允许用异常值惩罚因子 c 进行不完全分类。c 越大, 错分样本越少, 分类间距变小, 泛化能力减弱; c 越小, 错分样本越大, 分类间距变大, 泛化能力增强。

核函数的类型选择 RBF, 原因有三点: RBF 核函数可以将一个样本映射到一个更高维的空间, 而且线性核函数是 RBF 的一个特例, 也就是说如果考虑使用 RBF, 那么就没有必要考虑线性核函数; 需要确定的参数较少, 核函数参数的多少直接影响函数的复杂程度; 对于某些参数, RBF 和其他核函数具有相似的性能。在 RBF 核函数中自带一个 gamma 参数, 表示核函数的半径, 隐含地决定了数据映射到新的特征空间后的分布。

SVMtrain 对训练数据集进行训练, 获得 SVM 模型。模型内容如下:

```
svm_type c_svc      %训练所采用的 SVM 类型, 此处为 C-SVC
kernel_type rbf     %训练采用的核函数类型, 此处为 RBF 核
gamma 0.0769231 %设置核函数中的 gamma 参数, 默认值为 1/k
nr_class 2          %分类时的类别数, 此处为两分类问题
total_sv 132        %支持向量的总个数
rho 0.424462        %决策函数中的常数项
label 1 0           %类别标签
nr_sv 64 68         %各类别标签对应的支持向量个数
SV                  %以下为支持向量
1 1:0.166667 2:1 3:-0.333333 4:-0.433962 5:-0.383562 6:-1 7:-1
   8:0.0687023 9:-1 10:-0.903226 11:-1 12:-1 13:1
0.5104832128985164 1:0.125 2:1 3:0.333333 4:-0.320755
   5:-0.406393 6:1 7:1 8:0.0839695 9:1 10:-0.806452 12:-0.333333
   13:0.5
```

④采用十折交叉验证选择最佳参数 c 与 g(c 为惩罚系数, g 为核函数中 gamma 参数)。

交叉验证是把训练样本平均分成 10 份, 每次拿出 9 份当作训练集, 剩下的一份当作测试集, 这样重复 10 次, 获得 10 次的平均交叉验证准确率, 以此寻找最佳的参数, 使得准确率最高。在 LibSVM 源码中每次只能验证一组参数的效果, 要寻求最优参数, 只能多次手动设置参数。

本实验对源码进行修改, 采用网格搜索方法自动寻求最优参数并返回。具体操作为自动获取一组参数, 进行十折交叉验证, 得到平均准确率, 以此重复, 最终找到对应最高准确率的那组参数。为了确定训练集的合适大小, 选取三个训练集分别进行训练。实验结果表明, 训练集为 1 000 时, 平均分类精度为 80%; 训练集为 2 000 和 3 000 时, 平均分类精度都在 91%左右。因此, 为了保证训练集的精简, 训练集大小选择 2 000, 平均分类精度达到最高(91%)时, c 为 32, g 为 0.0625。

⑤采用最佳参数 c 与 g 对训练集进行训练获取 SVM 模型。

使用 SVMtrain 函数训练模型, LibSVM 中不会保存训练模型, 每次预测都需要重新训练。本实验改进了源码, 将训练好的模型进行本地保存, 以方便下次使用。

⑥利用获取的模型进行预测。

使用训练好的模型进行测试。输入新的 X 值, 给出 SVM 预测出的 Y 值。

4 可行性验证

4.1 验证方法

分别与两种方法进行对比实验, 验证基于 URL 特征的 Hub 网页识别方法的可行性: 与传统的基于多特征启发式规则的网页分类方法对比; 与传统的基于内容特征的机器学习方法对比。该阶段没有选用与传统基于 URL 的简单规则识别方法进行对比, 是因为在曹桂峰[6]的研究中已经明确证明基于 URL 简单规则的效果明显差于基于多特征启发式规则的分类方法。

可行性主要从效率和效果两方面进行验证, 已有研究在提出传统方法时, 只给出了其效果数据, 没有效率方面的数据, 因此本文将两种验证方法根据原有步骤再次实现, 在达到原有实验效果的同时得到其效率数据。

4.2 验证方法实现

(1) 基于多特征启发式规则的网页分类方法

①预处理操作。通过一组正则表达式去除注释信息、Script 脚本和 CSS 样式信息。

②计算网页的特征值。该过程是进行网页分类的关键, 主要是计算经过归一化后的非链接字符数、标点符号数、文字链接比。

③计算支持度。通过求得的各项特征值计算该网页为主题型网页的综合支持度。

④将计算得到的支持度同阈值进行比较。如果支持度小于该阈值, 则输出网页的类型为 Hub 网页, 否则输出网页类

型为主题型。

在该验证方法实现过程中，阈值是通过实验的方法获取，实验中选取 500 个 Hub 网页，计算每个网页为主题型网页的综合支持度，发现其值都集中在 0.6 以下，其中大部分集中在–0.2 以下，因此确定了阈值大致范围，最终在该范围内进行逐一测试实验，找出最优阈值，使得实验准确率最高。

(2) 基于内容特征的机器学习方法

①HTML 解析。通过建立 DOM 树去掉与网页分类无关的 HTML 源码。HTML 解析步骤如下:

1)规范化 HTML 标签

由于有些网页中的 HTML 标签是错误的、丢失的，为了后续处理的方便，需要将错误的标签改正回来，将丢失的标签补全。

2)建立 DOM 树

由 HTML 中的标签建立一棵 DOM 树。

3)网页去噪

除去<style>、<script>、<applet>等标签所嵌的 HTML 源码，因为这些代码只与网页表现形式有关，而与网页内容无关。

4)信息提取

从网页中抽取信息，包括：网页深度、更新周期、锚文本文字数量、文本文字数量、含有 URL 个数、含有新 URL 个数。

②特征提取。通过观察与实验，找出在两种网页类型上存在差异的特征项，因此提取 8 个网页内容特征，分别为：网页深度、更新周期、锚文本文字数量、文本文字数量、锚文本文字与文本文字数量之比、URL 个数、新 URL 个数、新 URL 所占比率。

③训练分类。根据提取的特征值建立训练集与测试集，训练 SVM 分类模型。

5 结果与分析

5.1 评价指标

对识别效果主要从两个方面进行评价：识别效率与识别效果。其中，效率就是系统开销，包括耗费时间、内存使用率和 CPU 使用率；效果主要包括准确率和召回率。

准确率(Precision)，表示在分类过程中得到的网页测试集中，网页类别被正确标注的网页所占的比率，反映分类器分类的准确程度。召回率(Recall)，表示在分类过程中得到的网页测试集中，真正网页类别被正确标注在所有符合该类别的网页测试集中所占的比率，反映分类器查到相关网页的完备性。

准确率与召回率反映分类效果的两个方面是互补的，单纯的提高某一个，另外一个就会受到其影响。在实际应用中，需要综合考虑准确率与召回率，目前主要用 F1 值作为评价标准，反映准确率和召回率综合效果。

实验采用 Precision, Recall 和 F1 评价网页分类的效果, Precision, Recall 和 F1 计算方法如下所示。参数含义如表 1 所示。

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

表 1 参数解释

Hub 网页		专家判断	
		Yes	No
分类器判断	Yes	TP	FP
	No	FN	TN

5.2 实验数据与环境

在实验中，从50个中文网站采集网页，这50个网站大致分为：政府网站(10个)、教育网站(10个)、事业单位网站(10个)、公司网站(10个)和新闻网站(10个)。其中，由于新闻网站网页数量较大，因此每个新闻网站采集500个网页，而其他4种类型网站相对较小，因此采集300个网页。为了保证数据的多样性，同时保证实验数据的简单高效，需要对采集的网页进行筛选精简，减少数据冗余。从采集的网页中分别选取1 000个、2 000个和3 000个作为训练集，每个训练集从每种类型网站中平均选取，其中 Hub 网页与主题网页数量各占一半。构造三个训练集是为了确定训练集大小的合适取值。

在训练机器学习模型时，是在随机切分的测试数据上得到的交叉验证平均准确率，因此对已有算法就不能使用同样的测试数据，造成缺乏可比性。因此在实验中又标注了另外 30 个网站中的 1 000 个网页作为测试数据，其中包括 600 个 Hub 网页和 400 个主题网页。保证了与已有算法的可比性，同时也在一定程度上证明本文提出算法的稳定性。

实验环境使用 Win7 系统, CPU 为 Intel 双核, 内存为 2GB。

5.3　实验结果

本实验分别选用基于 URL 特征的 Hub 网页识别方法、基于多特征启发式规则的网页分类方法和基于内容特征的机器学习方法进行三次实验。

表 2 为采用基于 URL 特征的 Hub 网页识别方法得到的实验结果, 经计算得到 Precision 为 91.20%, Recall 为 86.33%, F1 为 88.70%。在训练样本上对模型进行十折交叉验证得到平均准确率为 91%。

表 2　基于 URL 特征的 Hub 网页识别

Hub 网页		专家判断	
		Yes	No
分类器判断	Yes	518	50
	No	82	350

表 3 为采用基于多特征启发式规则的网页分类方法得到的结果, 实验中阈值在–0.2 到 0.6 之间选取, 经多次实验发现阈值选择–0.1 时, 准确率最高, 经计算得到 Precision 为 86.63%, Recall 为 83.17%, F1 为 84.86%, 该结果已达到曹桂峰[6]所做实验的效果。

表 3　基于多特征启发式规则的网页分类

Hub 网页		专家判断	
		Yes	No
分类器判断	Yes	499	77
	No	101	323

表 4 为采用基于内容特征的机器学习方法得到的实验结果, 经计算得到 Precision 为 88.73%, Recall 为 90.50%, F1 为 89.61%, 该结果已达到文献[9]中实验的效果。

表 4　基于内容特征的识别

Hub 网页		专家判断	
		Yes	No
分类器判断	Yes	543	69
	No	57	331

表 5 是在三种方法具体实现时, 得到的各个方法运行过程中消耗的时间、内存使用情况和 CPU 使用情况。

5.4　分析与讨论

为证明基于 URL 特征的 Hub 网页识别方法的稳定性, 在训练阶段对该模型进行了十折交叉验证, 得到平均准确率为 91%, 用该模型对测试数据进行测试时, 得到准确率为 91.2%, 这两组数据没有明显差异, 由此可以证明该方法具有一般性与稳定性。实验结果对比如图 2 所示:

表 5　三种方法的系统开销数据

实验组别	方法	处理网页数	耗时/s	内存使用/MB	CPU 使用率
1	基于多特征启发式规则	1 000	79.6	112	51%
2	基于内容特征	1 000	87.5	128	59%
3	基于 URL 特征	1 000	21.3	36	17%

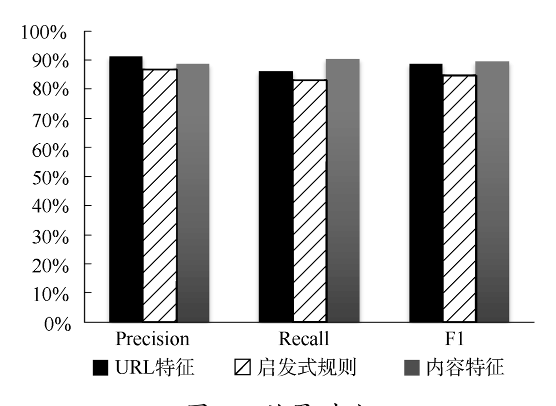


图 2　效果对比

基于 URL 特征的 Hub 网页识别方法优于基于多特征启发式规则的分类方法, 经分析原因为: 基于多特征启发式规则的分类方法缺乏灵活性, 不可能适用于所有网页; 在基于多特征启发式规则的分类方法中阈值的设定存在盲目性; 基于 URL 特征的 Hub 网页识别采用机器学习模型, 能发掘特征之间的内在联系, 具有很强的泛化能力。

基于 URL 特征的 Hub 网页识别方法与基于内容特征的机器学习方法在实验效果上没有太大差异, 因为两者采用了相同的识别方法, 只是选择的特征对象不同。基于 URL 特征的方法在准确率上优于基于内容特征的方法, 而在召回率上低于基于内容特征的方法, 原因是: Hub 网页所对应的 URL 特征明显, 如 URL 标题和 URL 长度较短、不包含日期等, 所以依据 URL 特征识别的准确率会相对较高。但 URL 存在很大随意性, 没有统一规范, 依据个人设定, 当不符合一般特性的 URL 出现时, 该方法很难识别, 所以召回率会相对较低; Hub 网页的内容存在一般特性, 如链接较多,

普通文本文字较少等，基本所有 Hub 网页都满足此特性，所以依据内容特征识别的召回率会很高。但有些主题网页也存在很多相关链接，其中内容文本也很短，所以依据内容特征识别的准确率会降低。综上所述，这两种方法在识别效果上差别很小，各有优点，但是在识别效率上存在明显差别。

如表 5 所示，基于 URL 特征的 Hub 网页识别方法在运行效率上有很大的优势，时间消耗少，大幅度降低了识别时间(70%)。因为 URL 本身相对较小，URL 特征提取就相对简单，但提取网页内容特征需要进行 HTML 解析，HTML 解析本身就是一项耗时的工作；内存与 CPU 占用较少，大约为传统方法的 60%，对采集系统影响小。因为在采集过程中本就会提取 URL，所以不会带来很大的额外开销，也不会影响采集系统的采集效率。综合以上原因，基于 URL 特征的 Hub 网页识别方法具有一定理论意义与实际应用价值，是一种行之有效的方法。

6 结　语

本文提出的基于 URL 特征的 Hub 网页识别技术，通过提取 URL 特征以训练机器学习模型，达到自动识别的目的。实验结果表明，该方法在达到传统方法识别效果的同时，能降低约60%的系统开销。但该方法存在一定局限性，因为 URL 本身具有一定随意性，当遇到 URL 特征不明显甚至完全相背的网站时，识别准确率会大幅度降低，此时需要结合网页内容特征去识别。因此，如何将基于内容特征的方法与基于 URL 特征的方法相结合以适应所有网站，是下一步研究的重点。

参考文献：

[1] 孟涛，闫宏飞，王继民. Web 网页信息变化的时间局部性规律及其验证[J]. 情报学报，2005，24(4)：398-406. (Meng Tao, Yan Hongfei, Wang Jimin. Characterizing Temporal Locality in Changes of Web Documents [J]. Journal of the China Society for Scientific and Technical Information, 2005, 24(4): 398-406.)

[2] 李晓明，闫宏飞，王继民. 搜索引擎——原理、技术与系统[M]. 北京:科学出版社，2005. (Li Xiaoming, Yan Hongfei, Wang Jimin. Search Engine: Theory, Technology and System [M]. Beijing: Science Press, 2005.)

[3] Cho J, Garcia-Molina H. The Evolution of the Web and Implications for an Incremental Crawler[C]. In: Proceedings of the 26th International Conference on Very Large Data Bases, 2002.

[4] Meng T, Yan H, Wang J, et al. The Evolution of Link-attributes for Pages and Its Implications on Web Crawling[C]. In: Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence, 2004.

[5] Ali R, Beg N M S. An Overview of Web Search Evaluation Methods [J]. Computers & Electrical Engineering,2011,37(6): 835-848.

[6] 曹桂峰. 搜索引擎中网页分类和网页净化的研究与实现[D]. 武汉：武汉理工大学, 2013. (Cao Guifeng. Design and Implement of Webpage Classify and Clean in Search Engine [D]. Wuhan: Wuhan University of Technology, 2013.)

[7] Zhang X, Zhou M, Geng G, et al. A Combined Feature Selection Method for Chinese Text Categorization [C]. In: Proceedings of the 2009 International Conference on Information Engineering and Computer Science, 2009.

[8] 谢光华. 中文网页自动分类的研究及其应用[D]. 大连：大连理工大学,2007. (Xie Guanghua. Research and Application of Chinese Web Page Automatic Classification [J]. Journal of Dalian University of Technology, 2007.)

[9] Wang R J, Wang D J. Web Information Acquisition by Personal Search Engine Based on SVM [J]. International Journal of Information Acquisition, 2005, 2(4): 345-352.

[10] 庞剑锋，卜东波，白硕. 基于向量空间模型的文本自动分类系统的研究与实现[J]. 计算机应用研究，2001，18(9): 23-26. (Pang Jianfeng, Bu Dongbo, Bai Shuo. Research and Implementation of Text Categorization System Based on VSM [J]. Application Research of Computers, 2001, 18(9): 23-26.)

[11] 李亮，刘万春，徐泉清，等. 一种基于支持向量机的专业中文网页分类器[J]. 计算机应用，2004，24(4)：58-61. (Li Liang, Liu Wanchun, Xu Quanqing, et al. A Professional Chinese Web Page Classifier Based on Support Vector Machine [J]. Computer Application, 2004, 24(4): 58-61.)

[12] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报，2000，26(1)：32-42. (Zhang Xuegong. Introduction to Statistical Learning Theory and Support Vector Machines [J]. Acta Automatica Sinica, 2000, 26(1): 32-42.)

[13] Chang C C, Lin C J. LIBSVM: A Library for Support Vector Machines [J]. Transactions on Intelligent Systems and

Technology, 2011, 2(3): Article No.27.

[14] Jiang J, Song X, Yu N, et al. Focus: Learning to Crawl Web Forums [J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(6): 1293-1306.

[15] Le A, Markopoulou A, Faloutsos M. PhishDef: URL Names Say It All [C]. In: Proceedings of the 30th IEEE International Conference on Computer Communications (INFOCOM), Shanghai, China. 2011.

作者贡献声明：

都云程：确定研究方向，提出研究思路，对论文提出修改意见；
张策：设计研究方案，进行实验过程设计及实验分析，起草论文；
梁然：采集基础数据；
张策，梁然：论文修改及最终版本修订。

收稿日期: 2015-06-25
收修改稿日期: 2015-08-13

A Study on Hub Page Recognition Using URL Features

Zhang Ce[1]　Du Yuncheng[1,2]　Liang Ran[2]
[1](Open Laboratory of TRS Software, Beijing Information Science and Technology University, Beijing 100085, China)
[2](Beijing TRS Information Technology Co. Ltd., Beijing 100101, China)

Abstract: [**Objective**] By building a simple data sample, the low efficiency as the problem of traditional recognition method is solved. [**Methods**] This method uses URL features as the basis of recognition, and uses Support Vector Machine (SVM) to recognize page type. [**Results**] The precision of this method is 91.2%, also in terms of efficiency performance, the method is increased by nearly 60%. [**Limitations**] When the URL feature is not obvious or even completely contrary, the recognition accuracy will be greatly reduced. [**Conclusions**] The experimental results show that the method has a great advantage in efficiency, and it will increase the efficiency of the collection system.

Keywords: URL features　Hub pages　SVM